El Corpus paralelo

Jos Hallebeek
Departamento de Español
Universidad de Nijmegen
Erasmusplein, 1
6500 HD Nijmegen (Países Bajos)
E-mail: J.Hallebeek@let.kun.nl

1. Corpus unilingües, bilingües, multilingües y paralelos

Un corpus de textos no contiene necesariamente textos en una sola lengua. Puede ser de dos (corpus bilingüe) o de más lenguas (corpus multilingüe). En tales casos los textos del corpus no son textos reunidos arbitrariamente, sino que están escogidos según idénticos criterios de selección en una y otra lengua. Por ejemplo, el *Aarhus Corpus of Danish, French and English* está compuesto de textos en tres lenguas que tratan todos del mismo tema: el derecho de contrato. No son traducciones de los mismos textos. Los textos son diferentes pero coinciden en la temática.

Cuando un corpus tiene los mismos textos en diferentes lenguas se habla de un *parallel corpus*: corpus paralelo. Hay que advertir que existe cierta confusión terminológica porque unos, por ejemplo Johansson & Hofland (1994: 25), Schmied & Schäffler (1996: 41) consideran que un corpus paralelo está compuesto de textos originales seleccionados según los mismos criterios y en diferentes lenguas, como es el caso del *Aarhus Corpus*. Si se trata de textos originales con su traducción hablan de un corpus de traducción.[1] Siguiendo la terminología de Baker (1995: 230) y McEnery (1996: 58), que parece ser la aceptada comúnmente ahora, denominaremos un corpus que consiste de textos originales en una lengua con su traducción en otra un corpus *paralelo*. Un corpus con textos en dos o más lenguas seleccionados según los mismos criterios lo llamamos un corpus bilingüe o multilingüe. Un ejemplo de un corpus paralelo es el *Hansard Corpus*, que contiene una colección de actas del parlamento canadiense tanto en versión inglesa como en la francesa. El corpus paralelo se usa para la construcción automática de léxicos y para la investigación sobre la traducción. Para poder manejar estos corpus es necesario que las frases y las palabras que son traducciones mutuas sean alineadas, es decir puestas unas al lado de otras.

El *Hansard Corpus* no es el único ejemplo de un corpus paralelo. Hay un corpus compuesto de manuales técnicos de IBM escritos en francés y en inglés. Luego en el proyecto

europeo CRATER (*Corpus Resources and Terminology Extraction*) se utiliza un corpus en tres lenguas: inglés, francés y español, del terreno de las telecomunicaciones. Se llama el ITU (*International Telecommunications Union*) corpus. El objetivo principal de CRATER es la confección automática de léxicos bilingües. Otros ejemplos de corpus que son paralelos en mayor o menor grado son:

- el corpus paralelo inglés - noruego (universidad de Oslo) (Johansson & Ebeling 1996);
- el corpus paralelo inglés - sueco (universidad de Lund) (M. Johansson 1996);
- el Chemnitz corpus paralelo/de traducción inglés - alemán (Schmied & Schäffler 1996);
- el corpus paralelo castellano - euskara (Abaitua, Casilla & Martínez 1997);
- el GILLBT Corpus de lenguas africanas;
- la ATR Dialogue Database, japonés - inglés;
- la Leiden-Jerusalem Armenian Database, armenio, griego, árabe y sirio (Baker 1995: 232).

Los cuatro primeros son corpora recientes sólo parcialmente acabados; los otros tres son menos recientes y están citados en Baker (1995: 232).

Cuando en un corpus se combinan fragmentos de textos originales y textos traducidos, todos en la misma lengua, Baker (1995: 234) habla de un corpus comparable (*comparable corpus*). El diseño (los criterios de selección) de los dos grupos de textos tendrá que ser idéntico. Este tipo de corpus se usa para estudios sobre la traducción a fin de determinar características generales de textos traducidos en comparación con textos originales (es decir, no traducidos).

Los textos traducidos incluidos en un corpus necesitan información detallada sobre la persona del traductor: profesional o aficionado, traduce si o no a la lengua materna. Esto es para poder sacar conclusiones válidas con respecto a las traducciones en comparación con los textos originales.

En la universidad de Oslo Johansson & Hofland (1994 y Johansson & Ebeling 1996) están compilando un corpus de un millón de palabras que consiste de 34 parejas de texto en inglés y en noruego. Son fragmentos de 10.000 ó 15.000 palabras sacados de obras originales y de su traducción al inglés o al noruego. Es decir que se parte de originales en las dos lenguas . No dan los datos exactos de los textos incluidos en el corpus. Se limitan a mencionar dos novelas noruegas y dos inglesas, el texto del acuerdo económico europeo, y del acuerdo de Maastricht (el último no en su poder todavía), y posibles textos de la agencia noruega de prensa. Como el corpus en inglés no sólo contiene textos originales sino también textos traduci-

dos del noruego, se trata al mismo tiempo de un corpus comparable. Por otra parte los corpus de los originales en inglés y en noruego están diseñados según los mismos criterios de selección, y en este sentido la combinación de los dos constituye un corpus bilingüe. Los textos que forman parte del corpus de un millón de palabras estarán alineados. En la parte del corpus alineada hasta la publicación de Johansson & Ebeling (1996) se han puesto una al lado de otra las frases (los enunciados) de los textos, no las palabras sueltas. El equipo tiene otros textos en preparación para agregar al corpus en algún momento del futuro. Los textos están codificados según las normas TEI (*Text Encoding Initiative*). Puesto que las unidades de los textos estarán identificadas detalladamente y los originales con sus traducciones estarán alineados, un programa de búsqueda producirá parejas de enunciados (frases) para cualquier palabra o expresión en cada una de las dos lenguas.

El corpus paralelo inglés - sueco está proyectado para incluir una variedad de géneros de texto según el modelo del LOB y del Brown corpus. Incluye textos originales en las dos lengas con sus traducciones, en fragmentos de 10.000 a 15.000 palabras. El corpus está en vías de compilación.

El Chemnitz corpus de traducción inglés - alemán es un corpus paralelo de textos originales tanto en inglés como en alemán junto con sus traducciones. Tendrá las siguientes categorías de texto, mayormente en fragmentos de 20.000 palabras:

- publicaciones de la comunidad europea en terreno de la economía y de la política social;
- libros de texto científicos: historia, filosofía, ciencias políticas, económicas y naturales;
- folletos turísticos;
- publicaciones de la Embajada Británica de Bonn;
- literatura contemporánea.

El número total de palabras será de 1.500.000. Los textos literarios ocupan sólo una parte marginal del corpus y no forman parte del *core corpus*.

El corpus castellano - euskara se enmarca en el proyecto LEGEBIDUNA en el que colaboran representantes de las universidades de Deusto, Alcalá de Henares y de la Complutense. Está compuesto de boletines oficiales de las Diputaciones de Álava y Bizkaia y del Gobierno Vasco, con aproximadamente 7 millones de palabras en cada una de las lenguas. Uno de los objetivos principales del proyecto es la creación de 'memorias de traducción' mediante el reconocimiento de unidades de traducción, que no son oraciones enteras sino partes de ellas. Estas unidades de traducción son sensibles al contexto, en el sentido de que se

identifican los diferentes registros del lenguaje (general o administrativo) al que pertenecen. El etiquetado del corpus se realiza en la línea de las propuestas TEI y MULTEXT.

2. Procedimientos utilizados en la alineación de palabras y frases[2]

Para llegar a alinear automáticamente las oraciones y las palabras de textos en lenguas diferentes se utilizan en primer lugar métodos puramente estadísticos. Así es por ejemplo el método Gale & Church (1993), desarrollado con ayuda del *Hansard Corpus*, que se basa en dos criterios:

a. oraciones más largas en una lengua se traducen en secuencias más largas en otra;
b. unos tipos de alineamiento se hallan con más frecuencia que otros: es más probable que una frase en una lengua se traduce también por una frase en la otra.

De modo que una frase traducida por dos, o dos frases traducidas por una son analogías menos frecuentes. En el proyecto CRATER se ha combinado el método de Gale & Church con otros sobre los que informan detalladamente McEnery & Oakes (1996). En un modelo probabilístico se combina la relación entre el número de caracteres de la frase en una lengua y en la otra, con el número de frases, dando *penaltys* para la falta de probabilidad. Si una frase en una lengua se traduce por una frase en la otra no se recibe un penalty, pero si una se traduce por dos el *penalty* es 230, etcétera. La corrección obtenida en los resultados de la alineación entre dos textos varía según la pareja de textos y según el tema del texto. Se llega a un 98% de corrección en los textos inglés y francés del ITU corpus. En textos periodísticos en inglés y en chino el porcentaje de corrección es de 54,5%.

Para mejorar los sistemas de alineamiento se han ido introduciendo en el reconocimiento de traducciones los llamados *cognates*, que no son sólo palabras sino más bien signos comunes a las dos lenguas. Se llaman *anchor points* ('puntos ancla'), o sea puntos de correspondencia conocida entre dos lenguas. Entre inglés y francés, por ejemplo, tenemos los siguientes: los signos de interrogación y de exclamación; las palabras que tienen cierto número de caracteres idéntico al principio de la palabra (*tax* y *taxe*); los nombres propios; las expresiones numéricas; las paréntesis. Se están llevando a cabo experimentos para comprobar la similitud entre dos palabras de diferentes lenguas, que no coinciden por completo. Hasta qué punto coinciden los caracteres y cuántas modificaciones son necesarias para convertir una palabra de una lengua en su equivalente en otra: por ejemplo, *couleur* y *colour* necesitan 2 cambios.

En el corpus paralelo inglés - noruego el procedimiento del alineamiento de los textos parte de las correspondencias en palabras ancla en los enunciados de ambos textos (el original y su traducción). Las palabras ancla forman en su sistema un léxico bilingüe compuesto según el criterio de que sean palabras de uso frecuente que tengan equivalentes directos en las dos lenguas. Están en esta lista de palabras ancla: palabras funcionales, palabras de contenido: nombres de los días, meses, adjetivos y nombres de uso frecuente (en total 850 entradas en Johansson & Hofland (1994: 29-32). También se utilizan partes de palabras (raíces): *open* está por *open, opens, opened, openly, openness,* etc. y nombres propios sacados del texto original de forma automatizada. Las parejas de frases que tienen el mayor número de palabras ancla en común con más probabilidad son original y traducción. Se incluye en el programa una consideración del tamaño del enunciado en número de palabras y número de caracteres. Con respecto a esto último conviene observar que en una lengua con pocas palabras compuestas (como el español) comparada con otra con muchas palabras compuestas (como el holandés) no coinciden la cantidad de palabras de un texto pero sí se acerca la cantidad de caracteres. Una palabra compuesta tiene más o menos el mismo número de caracteres que dos o más palabras simples que representan el mismo concepto. A base del número de caracteres combinado con el número de palabras ancla el programa permite también concluir que un solo enunciado de la lengua fuente se convierte en dos enunciados de la lengua objeto o al revés. La combinación ideal del *anchor score* y la cantidad de caracteres no la han encontrado todavía. En enunciados con pocas palabras ancla la cantidad de caracteres tiene un peso más alto. Hasta hora el alineamiento se limita a enunciados, es decir que no se alinean palabras.

3. Aplicaciones de corpus en la investigación sobre la traducción[3]

Los corpus paralelos bilingües se usan en la traducción automática en sistemas de traducción estadística basada en alineamiento léxico y la posición de palabras. Suelen tomarse en consideración no frases enteras sino secuencias de tres palabras, para las que se busca la equivalencia en otra lengua. Los corpus son una fuente de datos directa para las máquinas. A base del principio de la analogía sacan del corpus ejemplos típicos de frases o partes de ella para llegar a realizar la traducción de textos no traducidos todavía. Las últimas tendencias en los sistemas de traducción automática se alejan cada vez más de análisis sintácticos y semánticos completos utilizando gramáticas de reglas formales para ir basándose en datos de uso de la lengua viva.

En *Example-Based Machine Translation (EBMT)* la traducción se realiza según el siguiente procedimiento. Se dispone de un corpus bilingüe de textos traducidos y alineados. Al ofrecerse una frase para traducir, el programa busca en el corpus si ya está la misma frase alineada con su traducción en la otra lengua. La probabilidad de encontrar la misma frase ya traducida no es muy alta, a no ser que se disponga de un corpus de muchos millones de frases traducidas y alineadas. Por esto, también se ha ideado un sistema que no busca la traducción de la frase entera sino de partes de ella. Usando gramáticas sintácticas, tanto en la lengua fuente como en la lengua objeto se llega a dividir las frases en sus constituyentes funcionales. Luego se buscan las equivalencias de esas partes en una y la otra lengua. La ventaja de este sistema es que las frases completas no necesitan ser idénticas y que se pueden aprovechar partes de diferentes frases. El sistema no ha sido probado todavía.

De hecho, en la investigación sobre la traducción se utiliza toda clase de corpus: monolingües, bilingües, multilingües, paralelos (Baker 1995). Se estudia la relación entre el texto fuente en lengua A y su traducción a la lengua B. Pero también es interesante ver cómo en la lengua A se diferencian textos originales de textos traducidos de otras lenguas. En otras palabras el objeto de estudio es en el último caso el texto mismo. Lo que se hace es investigar cuáles son las características de textos traducidos en sí sin considerarlos sólo en relación con los originales. Ese tipo de corpus monolingüe estará compuesto por textos originales y textos traducidos en la misma lengua. Nos permite investigar las diferencias entre ambos tipos de textos. El corpus multilingüe con textos en diferentes lenguas seleccionados de acuerdo con los mismos criterios nos puede informar sobre las maneras en que se produce texto en esas lenguas y sobre las posibilidades (o la falta de ellas) de expresar las mismas cosas en distintas lenguas. El corpus paralelo que contiene textos originales con su traducción en otra u otras lenguas se utiliza en la formación de traductores y para mejorar los resultados de sistemas de traducción automática. En el primer caso, su contribución esencial es que contiene evidencia de cómo un traductor resuelve problemas que se ofrecen en la práctica de la traducción. Al comparar un corpus de textos originales sobre determinado tema con textos traducidos sobre el mismo tema y en la misma lengua pueden comprobarse ciertas propiedades características. Por ejemplo, que el número de tipos de palabras (*types*) es inferior en las traducciones, que la relación palabras gramaticales vs. palabras léxicas es más baja en los textos traducidos, o que determinadas estructuras oracionales son más típicas para una u otra clase.

En Baker (1993) la autora concluye que en los estudios de traducción se ha llegado al punto en que la disciplina está lista para pasar a la aplicación de las técnicas y la metodología

de la lingüística de corpus. Los estudios de la traducción van adquiriendo cada vez más el carácter de una ciencia teórica en lugar de ciencia aplicada. El empleo de grandes bases de datos dará más posibilidades para llegar a formular generalizaciones superando el nivel de los estudios fragmentarios tan corrientes hasta ahora. La utilización de largos corpus de textos traducidos permite pasar de la investigación sobre la significación (equivalencia entre original y su traducción) a la caracterización de la lengua de las traducciones. Se prestará más atención al establecimiento de normas para textos traducidos, una normativa que ocupa una posición intermedia entre la competencia y la actuación. La atención irá dirigida hacia la descripción del texto objeto (o sea la traducción) definiéndose así una rama descriptiva de la traductología. Johansson & Ebeling (1996: 4) describen así los diferentes tipos de estudios que permite llevar a cabo el corpus paralelo:

a. estudios contrastivos basados en textos paralelos en inglés y en noruego;
b. estudios contrastivos basados en textos originales y su traducción en las dos lenguas;
c. varios tipos de estudios de traducción: del inglés al noruego y vice versa; comparación entre textos originales y textos traducidos en la misma lengua; y caracteres generales de textos traducidos a base de las traducciones en las dos lenguas.

4. Estudios realizados

Hasta el momento se ha realizado un estudio sobre las interrogaciones en inglés y en noruego en Wikberg (1996) utilizando el corpus paralelo. Wikberg comparó textos procedentes de novelas: originales en inglés con su traducción al noruego, así como textos originales en noruego con su traducción al inglés. Sacó del corpus las oraciones de 7 a 10 palabras para obtener oraciones finitas y no fragmentos de oraciones. El signo de interrogación resultó ser muy útil para encontrar las interrogaciones en el texto original y traducido. Oraciones interrogativas no son siempre simples preguntas sino que expresan también ruegos, sugerencias, etc. Ha encontrado contrastes formales, semánticos y pragmáticos entre las dos lenguas. A veces una interrogación en una lengua se convierte en una oración declarativa en la otra. Además el autor ha podido establecer diferencias en la distribución de estructuras sintácticas equivalentes en los textos originales y sus traducciones.

Aprovechando el corpus paralelo inglés - sueco, M. Johansson (1996) estudia el fenómeno de *fronting* en sueco y en inglés partiendo de tres obras suecas y tres obras inglesas con sus correspondientes traducciones.[4] Los seis textos representan tres géneros: ficción

general, novela policíaca y autobiografía. Para Johansson *fronting* refiere a la colocación de cualquier elemento delante del sujeto (en una oración declarativa) en inglés y delante del verbo + sujeto en sueco, con excepción de conjunciones, pronombres relativos e interrogativos, interjecciones como *well, oh*, etc. Las diferencias y coincidencias existentes entre ambas lenguas en cuanto al fenómeno en cuestión aparece con más claridad en la comparación de textos originales con su traducción. En inglés el fenómeno de *fronting* es mucho menos frecuente que en sueco en el que el objeto directo o un complemento adverbial aparece muchas veces al principio de la oración. Esto ya se sabía por intuición. Ahora se ve confirmada esta intuición en la comparación de textos concretos. Johansson estudia detalladamente las clases de adverbiales y de otras funciones sintácticas que se hallan en posición inicial antepuestas al sujeto o al verbo + sujeto. En realidad hay cuatro posibilidades:

- un elemento ocupa la misma posición en sueco y en inglés;
- un elemento cambia de posición;
- se usa una estructura diferente (por ejemplo, un adverbio es sustituido por una proposición subordinada);
- se suprime el elemento en la traducción.

Utilizando el Chemnitz corpus de traducción inglés - alemán Schmied & Schäffler (1996) dedican atención al fenómeno de lo que en inglés se llama *translationese*: hablando en términos generales refiere este fenómeno a desviaciones del uso corriente de la lengua presentes en textos traducidos y causadas por influencia del texto original (fuente). Las desviaciones pueden ser simplemente faltas contra el sistema de la lengua destino (por ejemplo no usar en traducción española un subjuntivo porque no se usa en la lengua fuente). También puede que la traducción vaya en contra de la norma de uso establecida en determinada lengua (por ejemplo, usar en traducción holandesa una proposición *cleft* (hendida) que en español se usa con frecuencia pero que en holandés no es corriente aunque tampoco es imposible (o sea un error gramatical). Existen también características universales de traducción que no se deben a influencias de determinada lengua fuente, como la tendencia a simplificar oraciones, a desambiguar elementos, a evitar repeticiones (Baker 1993: 243). Hasta ahora existen pocos estudios sobre la norma de uso de lenguas específicas, sea inglés, alemán o español. Sólo al saber cómo es la norma podemos constatar desviaciones de la misma. El corpus *comparable* puede servir para llegar a definir esa norma. Si hay algún conocimiento suelto, está basado en estudios contrastivos o intuición sobre propiedades de lenguas. Así sabemos que en inglés se usan verbos modales (*may*, etc.) donde el alemán o el holandés emplean adverbios modales.

En español el empleo de los pronombres personales tónicos de sujeto es mucho más limitado que en inglés, alemán u holandés. ¿Es un fenómeno que pertenece a la norma y al sistema? Hay muchos factores que juntos juegan un papel en el análisis y la comparación de textos traducidos: gramaticalidad, aceptabilidad (sí o no según la norma), contexto sociocultural, interpretación correcta del significado. En su artículo Schmied & Schäffler estudian si es cierta la afirmación de que el inglés es menos directo que el alemán. Tratan de encontrar diferencias cuantitativas y cualitativas en los elementos y estructuras que expresan lo tentativo y lo indirecto: adverbios y verbos modales, construcciones impersonales, comentarios parentéticos, etc. Tomando en cuenta textos originales en inglés y en alemán así como traducciones de los textos ingleses al alemán hacen un análisis pormenorizado de los elementos indicados. Ese análisis es complicado y se corre el riesgo de incurrir en interpretaciones subjetivas.

5. Hacia un corpus paralelo holandés - español

Viendo las experiencias y los resultados provisionales obtenidos en la confección y el análisis de corpus paralelos nos parece interesante empezar un experimento con un corpus de este tipo que contiene textos originales en holandés y en español acompañados de sus traducciones a las respectivas lenguas. Una vez alineados los textos las aplicaciones para la investigación y la enseñanza de idiomas son múltiples. Por lo que respecta a esto último pensamos sobre todo en la formación de traductores y en los estudios de gramática descriptiva y contrastiva de las dos lenguas.

No abundan las traducciones de textos no literarios de índole general entre el español y el holandés. Nos referimos a manuales de historia, ciencias sociales, antropología, etc. Por otra parte, los textos que suelen producir las oficinas de la Unión Europea, muchas veces disponibles en las lenguas de la comunidad, son de carácter técnico y van destinados a especialistas en agricultura, política, economía, etc. Esto los hace poco interesantes para alumnos universitarios cuyos conocimientos de la segunda lengua que estudian no permite aún el manejo de textos de una complicación léxica de este tipo.

Se ha demostrado (Hulst 1996) que las traducciones de folletos turísticos no cumplen siempre con las exigencias de calidad que se les debería imponer. Aunque sí es cierto que representan un tipo de textos que en sí es interesante analizar por su carácter persuasivo.

Antes de decidir por un corpus extenso que contenga diversas variedades de géneros de textos, optamos por la formación de un corpus de tamaño limitado compuesto de textos

literarios tomados de novelas de autores contemporáneos. En general, actualmente la calidad de las traducciones de este tipo de obras es bastante buena. La fantasía y el dominio del lenguaje de los novelistas se reflejan en los textos traducidos, interesantes desde el punto de vista del estilo y de los recursos sintácticos, morfológicos y léxicos empleados.

Nos proponemos por lo tanto desarrollar un corpus paralelo de 32 fragmentos de 20.000 palabras cada uno, procedentes de 8 novelas originales españolas y 8 novelas originales holandesas, con sus correspondientes traducciones. Este corpus tendrá el carácter de un corpus piloto del volumen de unas 640.000 palabras en total. Sobre el diseño y el desarrollo de la creación del corpus paralelo español - holandés esperamos informar en una futura contribución.

Bibliografía

Abaitua Odriozola, J.K., A. Casillas Rubio & R. Martínez Unanue (1997): "Segementación de corpus paralelos para memorias de traducción". En *Procesamiento del Lenguaje Natural*, Revista núm.21 (1977), 17-30

Baker, Mona (1993): "Corpus linguistics and translation studies. Implications and applications". En Baker *et al.* (1993: 233-250)

Baker, Mona (1995): "Corpora in translation studies. An overview and some suggestions for future research". En *Target*, 7.2, 223-243

Baker, Mona, Gill Francis & Elena Tognini-Bonelli (eds)(1993): *Text and technology. In Honour of John Sinclair*. John Benjamins Publishing Company, Philadelphia/Amsterdam

Brown, P., J. Cocke, S. della Pietra, V. della Pietra, F. Jelinek, J. Laffarty, R. Mercer & P. Roosin (1990): "A statistical approach to machine translation En *Computational Linguistics*, 16(2), 79-85

Brown, P., V. della Pietra, S. della Pietra & R. Mercer (1993): "The mathematics of statistical machine translation: parameter estimation". En *Computational Linguistics*, 19(2), 263-301

Fries, Udo, Gunnel Tottie & Peter Schneider (eds)(1994): *Creating and using English Language Corpora*. Rodopi, Amsterdam - Atlanta, GA

Gale, W.A. & K.W. Church (1993): "A program for aligning sentences in bilingual corpora". In *Computational Linguistics*, 19(1), 75-102

Hulst, J. (1996): "Met een dagkaart van Amsterdam naar Madrid? Op weg naar een functioneel model voor vertaalkritiek". En *7e Symposium Spaans in Onderwijs, Onderzoek en Bedrijfsleven*, 1996, 17-33

Johansson, Mats (1996): "Fronting in English and Swedish: A text-based contrastive analysis". En Percy *et al.* (1996: 29-39)

Johansson, Stig & Jarle Ebeling (1996): "Exploring the English-Norwegian Parallel Corpus". En Percy *et al.* (1996: 3-15)

Johansson, Stig & Knut Hofland (1994): "Towards an English-Norwegian parallel corpus". En Fries *et al.* (1994: 25-37)

McEnery, Tony & Michael Oakes (1996): "Sentence and word alignment in the CRATER Project". En Thomas *et al.* (1996: 211-231)

Sadler, V. (1989): *Working with Analogical Semantics*. Foris, Dordrecht

Sánchez León, F. (1995): "Desarrollo de un etiquetador morfosintáctico para el español". En *Procesamiento del Lenguaje Natural*, Revista núm. 17 (septiembre de 1995), 14-28

Schmied, Josef & Hildegard Schäffler (1996): "Approaching translationese through parallel and translation corpora". En Percy *et al.* (1996: 41-56)

Thomas, Jenny & Mick Short (eds) (1996): *Using Corpora for Language Research*. Studies in Honour of Geoffrey Leech

Tsujii, J., S. Ananiadou, J. Carroll & S. Sekine (1991): *Methodologies for the Development of Sublanguage MT System II*, CCL UMIST Report No. 91/11

Wikberg, Kay (1996): "Questions in English and Norwegian: Evidence from the English-Norwegian Parallel Corpus". En Percy *et al.* (1996: 17-28)

Wilson, A. & A. McEnery (eds) (1994): *Corpora in Language Education and Research: A Selection of Papers from Talc94*, Unit for Computer Research on the English Language, Technical Papers 4 (special issue) Lancaster University

Zanettin, F. (1994): "Parallel words: designing a bilingual database for translator activities". En Wilson *et al.* (1994: 99-111)

---

1. Pero en el mismo artículo Johansson & Hofland llaman un corpus paralelo al corpus que contiene tanto textos originales en inglés y en noruego, como traducciones de textos del inglés al noruego y vice versa.

2. Ver McEnery & Oakes (1996)

3. Ver Brown *et al.* (1990, 1993), Hallebeek (1994), Sadler (1989), Tsujii *et al.* (1991)

4. En la introducción a su artículo M. Johansson (1996: 30) habla de dos veces tres obras (tres originales suecos y tres originales ingleses) pero en las referencias (1996: 38-39) sólo menciona los tres originales suecos con sus traducciones.